DOCUMENT RESUME

ED 268 132 TM 850 730

AUTHOR	Butler, Ronald W.
TITLE	Likelihood Methods for Adaptive Filtering and Smoothing. Technical Report #455.
INSTITUTION	Minnesota Univ., Minneapolis. School of Statistics.
PUB DATE	Oct 85
NOTE	32p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Algorithms; *Estimation (Mathematics); *Mathematical Models; *Maximum Likelihood Statistics; Monte Carlo Methods; *Prediction; Research Methodology; *Scoring; Simulation; Statistical Studies
IDENTIFIERS	Adaptive Filtering; EM Algorithm; Iterative Methods; Kalman Filtering; *Linear Models; Smoothing Methods

ABSTRACT

The dynamic linear model or Kalman filtering model provides a useful methodology for predicting the past, present, and future states of a dynamic system, such as an object in motion or an economic or social indicator that is changing systematically with time. Recursive likelihood methods for adaptive Kalman filtering and smoothing are developed. Unknown observational and systematic covariances in the dynamic linear model are estimated recursively using both the EM algorithm and the method of scoring. Such estimates when used in combination with the Kalman filter and fixed interval smoother allow for adaptive filtering and smoothing of state space predictors. In order to assess the performance of the EM and scoring methods for estimation, a small Monte Carlo simulation is performed. Two different types of adaptive filters constructed from the scoring method are described: (1) a fast filter that does not require storage of past data; and (2) a slower filter, that requires past data storage, and is based on the computation of a new maximum likelihood estimate. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED268132

Likelihood Methods for Adaptive Filtering and Smoothing

by

Ronald W. Butler

University of Minnesota
School of Statistics

Technical Report #455

UNIVERSITY OF MINNESOTA

SCHOOL OF STATISTICS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. W. Butler

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 850 730

Likelihood Methods for Adaptive Filtering and Smoothing

by

Ronald W. Butler

University of Minnesota
School of Statistics

Technical Report #455

October 1985

University of Minnesota
School of Statistics
Department of Applied Statistics
St. Paul, Minnesota 55108

TM 850.730

Summary

Recursive likelihood methods for adaptive Kalman filtering and smoothing are developed. Unknown observational and systematic covariances in the dynamic linear model are estimated recursively using both the EM algorithm and the method of scoring. Such estimates when used in combination with the Kalman filter and fixed interval smoother allow for adaptive filtering and smoothing of state space predictors.

Key words: Dynamic linear model, Kalman filtering, fixed-interval smoothing, adaptive filtering, conditional predictive likelihood, EM algorithm, scoring, studentized innovations, outliers.

1. Introduction

The dynamic linear model or Kalman filtering model provides a useful methodology for predicting the past, present, and future states of a dynamic system, such as an object in motion or an economic or social indicator that is changing systematically with time. The unknown pathway or state of the system is $z_0, z_1, \ldots$, an unknown sequence of r-dimensional random vectors. For example, if an object is in motion then z_t might be the location-velocity-acceleration vector at time t. Inference above z_t is possible by observing measurement x_t at time t where

$$x_t = A_t z_t + \epsilon_t \quad \text{for } t = 1, \ldots, T. \tag{1.1}$$

Equality (1.1) is the "observational equation" where $\{A_t\}$ derive from the nature (or geometry) of the observations, and measurement errors $\epsilon_1, \ldots, \epsilon_T$ are i.i.d. multivariate normal with mean 0 and unknown covariance Ω, *i.e.*, $MVN_k(0, \Omega)$. "Systematic equation"

$$z_t = B_t z_{t-1} + f_t \quad \text{for } t = 1, \ldots, T, \ldots \tag{1.2}$$

describes the dynamic change of the system where $f_1, \ldots, f_T, \ldots$ are i.i.d. $MVN_r(0, \Xi)$ errors that are independent of $\{\epsilon_t\}$. The $\{B_t\}$ are, for example, determined by the physical laws of motion with the local linearity of (1.2) sufficing for this and most other practical applications. The unknown parameter in this model is $(\Omega, \Xi) = \theta$ while the $\{z_t\}$ are random effects or unobserved data. The observed data is $(x_1, \ldots, x_T) = x$ and the unknown past and present states of the system are represented as $(z_0, \ldots, z_T) = z$. The generic symbol $f(\cdot \mid \cdot)$ will be used to specify the conditional density of the first argument given the second.

Based on observed x, section one considers the prediction of $z_0, \ldots, z_T$, the so-called "smoothing" problem, from a conditional predictive likelihood viewpoint (Butler (1985)) and a conjugate Bayesian perspective. Both methods lead to iteratively-computed modal predictors similar in spirit to those of Lindley and Smith (1972) which, however, converge to a trivial mode in which either (1.1) or (1.2) has error of less than full rank.

Our lack of success in the direct implementation of predictive methods leads to the consideration of an estimative approach based on $f(z \mid x, \hat{\theta})$ where $\hat{\theta}$ is the maximum likelihood estimate (MLE) of θ. Recursive formulae for computing $\hat{\theta}$ based on the EM algorithm and the method of scoring are presented in sections three and four. The former relies on the fixed-interval smoothing equations as well as other results which are derived from a Bayesian viewpoint. The latter is implemented by a derivation of the Fisher information for θ.

Section five includes a small Monte Carlo simulation designed to assess the accuracy of $\hat{\theta}$ when computed from either the EM algorithm or the method of scoring. The predictive ability of the Kalman filter using the current value of $\hat{\theta}$, *i.e.*, the adaptive Kalman filter, is illustrated in the last section. Informal residual methods for detecting outliers in the observational equation are also considered.

2. Kalman Filtering and Predictive Smoothing

Likelihood prediction of z_{T+1} for known value of $\theta = (\Omega, \Xi)$ and unknown value of $z = (z_0, \ldots, z_T)$ is achieved by considering

$$f(z_{T+1} \mid x, \theta) = \int f(z_{T+1} \mid z_T, \Xi) f(x \mid z, \Omega) f(z \mid \Xi) dz \tag{2.1}$$

where $f(z_{T+1} \mid z_T, \Xi)$ is $MVN_r(B_{T+1} z_T, \Xi)$, $f(x \mid z, \Omega)$ is as specified conditionally in (1.1), and $f(z \mid \Xi)$ is as in (1.2). It is also necessary to assume $z_0 \sim MVN_r(\tilde{z}_0, \Sigma_0)$ with known values of $\tilde{z}_0$ and Σ_0 in order to initiate the filtering recursions. According to (1.1) and (1.2), z_{T+1} and x may be written exclusively in terms of $\{\epsilon_t\}$ and $\{f_t\}$ so that $z_T, x \mid \theta \sim MVN_{r+kT}$ and (2.1) is an MVN_r law. The mean and covariance of this law are conveniently derived from the Kalman filtering recursions. In particular,

$$E(z_{T+1} \mid x, \theta) = \tilde{z}_{T+1} = B_{T+1} \tilde{z}_T$$

and

$$V(z_{T+1} \mid x, \theta) = B_{T+1} \Sigma_T B'_{T+1} + \Xi = R_{T+1}$$

are the endproducts of recursions

$$\left.\begin{aligned} &\tilde{z}_t = B_t\tilde{z}_{t-1} + R_tA_t'(A_tR_tA_t' + \Omega)^{-1}(x_t - A_tB_t\tilde{z}_{t-1}) \\ &\text{where} \\ &R_t = B_t\Sigma_{t-1}B_t' + \Xi \\ &\text{and} \\ &\Sigma_t = R_t - R_tA_t'(A_tR_tA_t' + \Omega)^{-1}A_tR_t \end{aligned}\right\} \tag{2.4}$$

for $t = 1, \ldots, T$. The recursion in (2.4) from time $t-1$ to t may be calculated in "real time" and has been shown equivalent to the use of the Bayes theorem for sequential updating by Harrison and Stevens (1971, 1976). An elementary tutorial is found in Meinhold and Singpurwalla (1983).

When the aim is prediction of $z_0, \ldots, z_T$, a conditional predictive likelihood approach as in Butler (1985 (2.4)) can be used. The predictive likelihood of z given x is defined as

$$\mathcal{L}(z \mid x) = \frac{f(x, z \mid \theta)}{f(t(x,z) \mid \theta)|JJ'|^{\frac{1}{2}}}$$

where $t(x,z)$ is a minimal sufficient reduction of (x,z) for θ and $J = \partial t/\partial(x,z)$. Now $t(x,z) = (\hat{S}_\omega, \hat{S}_\xi)$ where

$$\begin{aligned} (T-k-1)\hat{\Omega} &= \sum_{t=1}^{T}(x_t - A_tz_t)^2 = \hat{S}_\omega \\ (T-r-1)\hat{\Xi} &= \sum_{t=1}^{T}(z_t - B_tz_{t-1})^2 = \hat{S}_\xi \end{aligned} \tag{2.5}$$

and the square of column vector x_t is x_tx_t', so that, with $\hat{S}_\omega \sim$ Wishart(T, Ω) and $\hat{S}_\xi \sim$ Wishart(T, Ξ), then

$$\mathcal{L}(z_0, \ldots, z_T \mid x) \propto |\hat{S}_\omega|^{-\frac{1}{2}(T-k-1)}|\hat{S}_\xi|^{-\frac{1}{2}(T-r-1)} \exp\{-\tfrac{1}{2}\operatorname{tr}\Sigma_0^{-1}(z_0 - \tilde{z}_0)^2\}/|JJ'|^{\frac{1}{2}}. \tag{2.6}$$

Modal predictors can be derived from (2.6) by differentiation. However, in order to interpret their behaviour, the "smoothed" predictors of z need to be defined.

Since x and z are linear combinations of $\{\epsilon_t\}$ and $\{f_t\}$, then $x, z \mid \theta \sim MVN$ and $z \mid x, \theta \sim MVN$. Therefore, for known θ and x, the joint modal predictor of $\{z_t\}$ is

$$\{E(z_t \mid x, \theta) = \hat{z}_t(\theta) = \hat{z}_t = (\hat{z}_{t1}, \ldots, \hat{z}_{tr})'\}.$$

This is called the "smoothed" predictor of z_t since it is calculated from data observed both before and after time t. Note that $\{\hat{z}_t\}$ satisfy the following $(T+1)r$ equations:

$$\begin{aligned} 0 = \frac{\partial(-2\ell n\, f(z \mid x, \theta))}{\partial z_{tj}} &= \mathrm{tr}\{\Omega^{-1}\partial\hat{S}_\omega/\partial z_{tj} + \Xi^{-1}\partial\hat{S}_\xi/\partial z_{tj}\} && \text{if } t \geq 1 \\ &= \mathrm{tr}\{\Xi^{-1}\partial\hat{S}_\xi/\partial z_{0j} + \Sigma_0^{-1}\partial(z_0 - \tilde{z}_0)^2/\partial z_{0j}\} && \text{if } t = 0 \end{aligned} \tag{2.7}$$

for $j = 1, \ldots, r$ where matrix differentiation is componentwise.

Differentiation of (2.6) using Mardia, Kent, and Bibby (1979, equation A.9.3) and the chain rule leads to an expression for $\partial(-2\ell n\, \mathcal{L})/\partial z_{tj}$ which agrees with the right side of (2.7) with $(\hat{\Omega}, \hat{\Xi})$ replacing (Ω, Ξ). This demonstrates that the modal likelihood predictors can be computed iteratively by first figuring the smoothed predictors for precalculated $(\hat{\Omega}, \hat{\Xi})$, recalculating $(\hat{\Omega}, \hat{\Xi})$ as in (2.5), and so on until a limit is achieved. Various numerical examples have demonstrated that this sequence of $\hat{z}$'s converges to a point at which either $|\hat{\Omega}| = 0$ or $|\hat{\Xi}| = 0$. (Note that if $k > r$ then $|\hat{\Omega}| > 0$ and only $|\hat{\Xi}| = 0$ is a limit point.) The iterations do converge to the mode of $\mathcal{L}$ which is the unbounded location where either the systematic or observational equation holds without $\hat{S}_\omega$ or $\hat{S}_\xi$ having full rank. Note, however, that when $k = r$ then z has a dimension greater than x so we are predicting more unknowns than we have knowns. This lack of full rank error when maximizing the predictive likelihood is analogous to the estimation of regression coefficients in Anova when so many independent variables are fit that there are no degrees of freedom left over for error.

A conjugate Bayesian computation of the predictive density $f(z \mid x)$ using independent inverted Wishart priors $\text{Wishart}^{-1}(\eta_1, \Omega_0^{-1})$ and $\text{Wishart}^{-1}(\eta_2, \Xi_0^{-1})$ for Ω and Ξ leads to

$$f(z \mid x) \propto |\Omega_0 + \hat{S}_\omega|^{-\frac{1}{2}(T+\eta_1)} |\Xi_0 + \hat{S}_\xi|^{-\frac{1}{2}(T+\eta_2)} \tag{2.8}$$

which is bounded for $\Omega_0 > 0 < \Xi_0$. Unfortunately the modal Bayesian predictors of (2.8) were the same as those from predictive likelihood in all our numerical examples. Thus both methods lead to trivial predictors which contrasts with the nontrivial predictors of Lindley and Smith (1972). Of course the dimension of z here is considerably higher than that of their random effects. This apparently is at the root of the problem since the modal predictors from the iterations can roam freely about in the z-space. When attention is restricted to subsets of z, then both methods become intractable.

3. The EM Algorithm

Maximum likelihood estimation of $\theta = (\Omega, \Xi)$ can be based on this algorithm by assuming $z = (z_0, \ldots, z_T)$ is missing data and

$$f(x \mid \theta) = \int f(x \mid z, \Omega) f(z \mid \Xi) dz$$

is the likelihood. The complete data sufficient statistics in (2.5) arise in the context of a regular exponential family so that the EM iterations are based on

$$\hat{\theta}^{(p+1)} = (\hat{\Omega}^{(p+1)}, \hat{\Xi}^{(p+1)}) = E((\hat{\Omega}, \hat{\Xi}) \mid x, \hat{\theta}^{(p)}) \tag{3.1}$$

as in (2.2) and (2.3) of Dempster *et al.* (1977) where T is now the divisor in the covariance computations of (2.5). The basic terms required in the computation of (3.1) are $E(z_t \mid x, \hat{\theta}^{(p)}) = \hat{z}_t(\hat{\theta}^{(p)}) = \hat{z}_t^{(p)}$ and $\mathrm{Var}(z_t \mid x, \hat{\theta}^{(p)}) = V_t^{(p)}$ for $t = 0, \ldots, T$, and

$$E((z_s - \hat{z}_s^{(p)})(z_t - \hat{z}_t^{(p)})' \mid x, \hat{\theta}^{(p)}) = V_{s,t}^{(p)}$$

for $|s - t| = 1$ where $V_{t,s}^{(p)} = (V_{s,t}^{(p)})'$. The relationship between (3.1) and these quantities is given by

$$E(\hat{\Omega} \mid x, \hat{\theta}^{(p)}) = T^{-1} \sum_{t=1}^{T} \{A_t V_t^{(p)} A_t' + (x_t - A_t \hat{z}_t^{(p)})^2\} \tag{3.2}$$

and

$$\begin{aligned} E(\hat{\Xi} \mid x, \hat{\theta}^{(p)}) = T^{-1} \sum_{t=1}^{T} \{ & V_t^{(p)} + (\hat{z}_t^{(p)} - B_t \hat{z}_{t-1}^{(p)})^2 + B_t V_{t-1}^{(p)} B_t' \\ & + V_{t,t-1}^{(p)} B_t' + B_t V_{t-1,t}^{(p)}\}. \end{aligned} \tag{3.3}$$

The quantities $\{\hat{z}_t^{(p)}\}$ and $\{V_t^{(p)}\}$ arise directly from the fixed interval smoothing recursions; however, the values of $\{V_{s,t}^{(p)}\}$ are not standard output. All these quantities are computed in Appendix 1 from a Bayesian viewpoint. The calculations rely on having run the Kalman filter with $\hat{\theta}^{(p)}$ forward in time and having saved the values $\{\tilde{z}_t, \Sigma_t, R_t\}$ whose dependence on $\hat{\theta}^{(p)}$ has been suppressed. Then $\{\hat{z}_t^{(p)}, V_t^{(p)}, V_{s,t}^{(p)}\}$ are calculated recursively in a backward time direction. These recursions are started by taking $\hat{z}_T^{(p)} = \tilde{z}_T$ and $V_T^{(p)} = \Sigma_T$. Then the imputed data is calculated as

$$\begin{aligned} \hat{z}_t^{(p)} &= \tilde{z}_t + S_t(\hat{z}_{t+1}^{(p)} - B_{t+1}\tilde{z}_t) \\ V_t^{(p)} &= \Sigma_t + S_t(V_{t+1}^{(p)} - R_{t+1})S_t' \\ V_{t+1,t}^{(p)} &= [(\hat{\Xi}^{(p)})^{-1} + (V_{t+1}^{(p)})^{-1} - R_{t+1}^{-1}]^{-1}(\hat{\Xi}^{(p)})^{-1}B_{t+1}V_t^{(p)} \end{aligned} \tag{3.4}$$

where

$$S_t = \Sigma_t B_{t+1}' R_{t+1}^{-1}$$

and R_{t+1} is given in (2.4) for $t = T-1, \ldots, 0$. The equations of (3.4) are the E-step when computed using current parameter estimates. Those of (3.2) and (3.3) comprise the M-step. The steps of this algorithm are summarized.

1. Take an initial guess $\hat{\theta}^{(0)} = (\hat{\Omega}^{(0)}, \hat{\Xi}^{(0)})$.
2. Run the Kalman filter of (2.4) forward in time using the current $\hat{\theta}$ and saving $\{\tilde{z}_t, \Sigma_t, R_t\}$.
3. Run the smoothing recursions of (3.4) backward in time using the input of step 2.
4. Compute an updated $\hat{\theta}$ as in (3.2) and (3.3) using the imputed data output of step 3.
5. Use a stopping rule to decide whether to return to step 2 or not.

4. The Method of Scoring

The iterative method of scoring (Kendall and Stuart (1973), §18.21) is based on the recurrence

$$\hat{\theta}^{(p+1)} = \hat{\theta}^{(p)} + I(\hat{\theta}^{(p)})^{-1}\dot{\ell}(\hat{\theta}^{(p)}) \tag{4.1}$$

where $\dot{\ell}(\theta) = \partial \ell n\, f(x \mid \theta)/\partial\theta$ is the score and $I(\theta) = -E(\partial^2 \ell n\, f(x \mid \theta)/\partial\theta\partial\theta')$ is the Fisher information. For notational convenience the $p = k(k+1)/2 + r(r+1)/2$ parameters of (Ω, Ξ) are identified as $(\theta_1, \ldots, \theta_p)' = \theta$ in some systematic way. Denoting $I_t = I_t(\theta)$ as the Fisher information at time t based on $x_1, \ldots, x_t$, then a recursive computation of I_t and I_t^{-1} in t follows from the derivation of I_t in Appendix 2. Let

$$f_i^t = \text{vec}(\partial G_t/\partial\theta_i) \quad \text{with } G_t = A_t R_t A_t' + \Omega = \text{Cov}(x_t \mid x_1, \ldots, x_{t-1})$$

where vec$(\cdot)$ is the operator stacking the columns of its matrix argument into a vector. Also define

$$\begin{aligned} F_t &= (f_1^t, \ldots, f_p^t), \\ h_i^t &= \text{vec}(G_t^{-1}\partial G_t/\partial\theta_i G_t^{-1}) \quad \text{and} \quad H_t = (h_1^t, \ldots, h_p^t), \\ \text{and} \quad M_t &= (\partial m_t/\partial\theta_1, \ldots, \partial m_t/\partial\theta_p) \quad \text{where} \\ m_t &= A_t B_t \tilde{x}_{t-1} = E(x_t \mid x_1, \ldots, x_{t-1}). \end{aligned} \tag{4.2}$$

Then

$$I_t = I_{t-1} + 1/2\, H_t' F_t + M_t' G_t^{-1} M_t \tag{4.3}$$

follows from (A2.9) with $I_0 = 0$. A recursion for finding I_t^{-1} from I_{t-1}^{-1} follows from a double application of (A1.2). Then

$$I_t^{-1} = \Psi_t^{-1} - \Psi_t^{-1} M_t'(G_t + M_t \Psi_t^{-1} M_t')^{-1} M_t \Psi_t^{-1} \tag{4.4}$$

$$\text{where} \quad \Psi_t^{-1} = I_{t-1}^{-1} - \tfrac{1}{2}\, I_{t-1}^{-1} H_t'(I + \tfrac{1}{2}\, F_t I_{t-1}^{-1} H_t')^{-1} F_t I_{t-1}^{-1} \tag{4.5}$$

and I is a $(k^2 \times k^2)$ identity matrix. Expressions (4.4) and (4.5) involve the inversion of $(k \times k)$ and $(k^2 \times k^2)$ matrices while the direct use of (4.3) requires a $(p \times p)$ inversion; therefore (4.4)-(4.5) are more computationally efficient when $2 \leq k \leq r$ and (4.3) is otherwise. Note that when $k < r$ then I_1 is not invertible since its rank is $k^2 + k < p$. Therefore a few iterations of (4.3) are required before (4.4) and (4.5) can be used. The components of matrices H_t, F_t, and M_t are also computed recursively and this is considered after discussing the score statistic.

A recursion for the score statistic follows from (A2.6). Denoting $\dot{\ell}_t = \dot{\ell}_t(\theta)$ as the score at time t, then

$$\dot{\ell}_t = \dot{\ell}_{t-1} + 1/2\, H_t' g_t + M_t' G_t^{-1}(x_t - m_t) \tag{4.6}$$

where

$$g_t = \text{vec}((x_t - m_t)^2 - G_t) \quad \text{and} \quad \dot{\ell}_0 = 0.$$

Expressions (4.3) and (4.6) reveal that $I_t^{-1}\dot{\ell}_t$ takes the form of a regression which weights the values $\dot{\ell}_{t-1}, g_t$, and $x_t - m_t$.

Appendix 2 develops the recursions required to figure H_t, F_t, and M_t. One should not let the complexity of these recursions overshadow the simplicity of the actual calculations.

1. Derivatives with respect to Ω.

(a) $\dot{G}_t = \partial G_t / \partial \omega_{ij} = \begin{cases} A_t \dot{R}_t A_t' + J_{ii} & \text{if } i = j \\ A_t \dot{R}_t A_t' + J_{ij} + J_{ji} & \text{if } i \neq j \end{cases}$

where J_{ij} is a $(k \times k)$ indicator matrix of component i, j and $\dot{R}_t = B_t \dot{\Sigma}_{t-1} B_t'$. Now $\dot{\Sigma}_{t-1}$ for $t \geq 2$ may be computed recursively from $\dot{R}_{t-1}$ using

$$\begin{aligned} \dot{\Sigma}_t = \dot{R}_t &- \dot{R}_t A_t' G_t^{-1} A_t R_t - R_t A_t' G_t^{-1} A_t \dot{R}_t \\ &+ R_t A_t' G_t^{-1}(A_t \dot{R}_t A_t' + \dot{\Omega}) G_t^{-1} A_t R_t \end{aligned} \tag{4.7}$$

where $\dot{\Omega} = J_{ii}$ if $i = j$ and $\dot{\Omega} = J_{ij} + J_{ji}$ if $i \neq j$.

(b) $\dot{m}_t = \partial m_t / \partial \omega_{ij} = A_t B_t \dot{\tilde{z}}_{t-1}$, where for $t \geq 1$,

$$\dot{\tilde{z}}_t = (I - R_t A_t' G_t^{-1} A_t)\{B_t \dot{\tilde{z}}_{t-1} + (\dot{R}_t A_t' - R_t A_t' \Omega^{-1} \dot{\Omega}) G_t^{-1}(x_t - A_t B_t \tilde{z}_{t-1})\}. \tag{4.8}$$

2. Derivatives with respect to Ξ.

(a) $\dot{G}_t = \partial G_t / \partial \xi_{ij} = A_t \dot{R}_t A_t'$ where

$$\dot{R}_t = \begin{cases} B_t \dot{\Sigma}_{t-1} B_t' + J_{ii} & \text{if } i = j \\ B_t \dot{\Sigma}_{t-1} B_t' + J_{ij} + J_{ji} & \text{if } i \neq j \end{cases}$$

and

$$\dot{\Sigma}_t = (I - R_t A_t' G_t^{-1} A_t)\dot{R}_t(I - A_t' G_t^{-1} A_t R_t) \quad \text{for } t \geq 1. \tag{4.9}$$

(b) $\dot{m}_t = \partial m_t/\partial \xi_{ij} = A_t B_t \dot{\tilde{z}}_{t-1}$ where for $t \geq 1$,

$$\dot{\tilde{z}}_t = (I - R_t A_t' G_t^{-1} A_t)\{B_t \dot{\tilde{z}}_{t-1} + \dot{R}_t A_t' G_t^{-1}(x_t - A_t B_t \dot{\tilde{z}}_{t-1})\}. \tag{4.10}$$

The steps to compute $I_T^{-1}(\theta)$ and $\dot{\ell}_T(\theta)$ for use in scoring are now summarized. These steps when implemented using the current $MLE\ \hat{\theta}^{(p)}$ allow $\hat{\theta}^{(p)}$ to be scored as in (4.1). An arbitrary step at time $t = 1, \ldots, T-1$ is considered:

1. Values for $I_{t-1}^{-1}, \dot{\ell}_{t-1}, R_{t-1}, \{\dot{R}_{t-1}\}, \tilde{z}_{t-1}$, and $\{\dot{\tilde{z}}_{t-1}\}$ have been saved from the previous step.
2. Include the information in x_t by computing
   (a) $\tilde{z}_t$ and R_t from the Kalman filter (2.4),
   (b) $\{\dot{R}_t\}$ and $\{\dot{\tilde{z}}_t\}$ from (4.7)-(4.10),
   (c) F_t, H_t, M_t, and g_t as in (4.2) and (4.6), and
   (d) I_t^{-1} and $\dot{\ell}_t$ as in (4.4)-(4.6).
3. Save the values for time t as in step 1 and continue on to include the information in x_{t+1}.

The starting values in these recursions are

$$I_0 = 0,\ \dot{\ell}_0 = 0,\ R_1 = B_1 \Sigma_0 B_1' + \Xi,\ \partial R_1/\partial \omega_{ij} = 0,$$

$$\partial R_1/\partial \xi_{ij} = \begin{cases} J_{ii} & \text{if } i = j \\ J_{ij} + J_{ji} & \text{if } i \neq j, \end{cases}$$

$$\tilde{z}_0 \quad \text{as specified, and} \quad \dot{\tilde{z}}_0 = 0.$$

As previously mentioned, the first few recursions must update I_t as in (4.3) until I_t is invertible.

5. Monte Carlo

In order to assess the performance of the EM and scoring methods for estimation, a small Monte Carlo simulation was performed. Taking $r = 1 = k$, $A_t = 1$, $B_t = (-1)^t/2$, $\Omega = 2$, and $\Xi = 1$, the Kalman filtering model of (1.1)-(1.2) was simulated 20 times using $T = 50$, 200, and 500 in order to evaluate the EM algorithm. Table 1 summarizes these results. Each of the 20 determinations of $\hat{\theta}$ in the trials is based on the best of six attempts to find the MLE starting at randomly generated values of Ω and Ξ. Best here is measured in terms of the likelihood value at the 200^{th} iteration. The starting values were $\hat{\Omega}^{(0)} \sim \chi^2_4$ and $\hat{\Xi}^{(0)} \sim \chi^2_2$ which were often quite far from 2 and 1.

The consistency of $\hat{\theta}$ as $T \to \infty$ is reflected in the decreasing error rates in Table 1 as T increases. The table also suggests that Ξ can be estimated with greater accuracy than Ω. This was unexpected since data derives directly from the observational equation with variance Ω whereas Ξ reflects differences in unknown state values and must be inferred somewhat indirectly. In many of the repetitions the EM iteration sequences were observed to make very slow progress in locating modal values. There was no indication of multiple local maxima in any of the trials.

The same model and sample sizes were used in assessing the scoring method using 50 repetitions. The conditions of this simulation differed only in terms of the greater number of repetitions (made possible because of less computation) and fewer iterations (100) per repetition. In several instances scoring iterations produced negative variance estimates so it was necessary to constrain $\hat{\theta}$ values to lie in $[.1, \infty)^2$.

Table 1. Summary Statistics for EM Algorithm
(Each entry is an average of 20 values)

$T =$		50	200	500
$\hat{\Omega}$:	mean	1.675	2.066	1.900
	MAE^*	.692	.483	.179
	$MSE^{\frac{1}{2}}$	.809	.572	.232
$\hat{\Xi}$:	mean	1.169	.918	.961
	MAE	.667	.416	.176
	$MSE^{\frac{1}{2}}$	.788	.436	.236

*MAE = mean absolute error about the true value and MSE = mean squared error.

Table 2. Summary Statistics from Scoring
(Each entry is an average of 50 values)

$T =$		50	200	500
$\hat{\Omega}$:	mean	2.150	2.043	2.111
	MAE	.752	.406	.250
	$(MSE)^{\frac{1}{2}}$	.919	.498	.324
$\hat{\Xi}$:	mean	1.095	1.093	1.060
	MAE	.594	.331	.254
	$(MSE)^{\frac{1}{2}}$	.741	.397	.302
$\hat{V}(\hat{\Omega})^*$:	mean	1.130	.256	.104
	$(\text{mean})^{\frac{1}{2}}$	1.063	.506	.322
	SD	.413	.0441	.0109
$\hat{V}(\hat{\Xi})^*$:	mean	.882	.204	.0806
	$(\text{mean})^{\frac{1}{2}}$	.939	.452	.284
	SD	.393	.0475	.0163
$\widehat{\text{Cov}}(\hat{\Omega}, \hat{\Xi})^*$:	mean	−.784	−.180	−.0720
	SD	.273	.0306	.0101
$\widehat{\text{Corr}}(\hat{\Omega}, \hat{\Xi})$:	mean	−.801	−.792	−.788
	SD	.0237	.0128	.00811

* These are the components of the inverse of the estimated Fisher information.

Table 2 summarizes these results and corroborates the findings of Table 1. The last four entries are variances, covariances, and correlations from the inverse of the estimated Fisher information which estimates the asymptotic covariance structure of $\hat{\theta}$. The table reveals estimation of Ξ to be asymptotically more efficient than for Ω under this model. The finite sample $MSE^{\frac{1}{2}}$ values of $\hat{\Omega}$ and $\hat{\Xi}$ are quite close to the estimated asymptotic standard deviations listed as $(\text{mean})^{\frac{1}{2}}$ suggesting that the limiting asymptotic normal distribution may closely approximate the finite sample distribution of $\hat{\theta}$. The normal probability plots in Figures 1 and 2 are fairly linear, also supporting this contention. Figures 3 and 4 are normal plots with each value standardized by its estimated asymptotic standard deviation. These plots are somewhat rough but overall linear. The large asymptotic correlation of $\hat{\Omega}$ and $\hat{\Xi}$ suggests a possible reason why the EM algorithm is so slow in converging. With T large the likelihood is approximately a bivariate normal with correlation $-.79$. Such a density has a long gradual ridge which could take many iterations for the EM algorithm to climb.

Figure 1

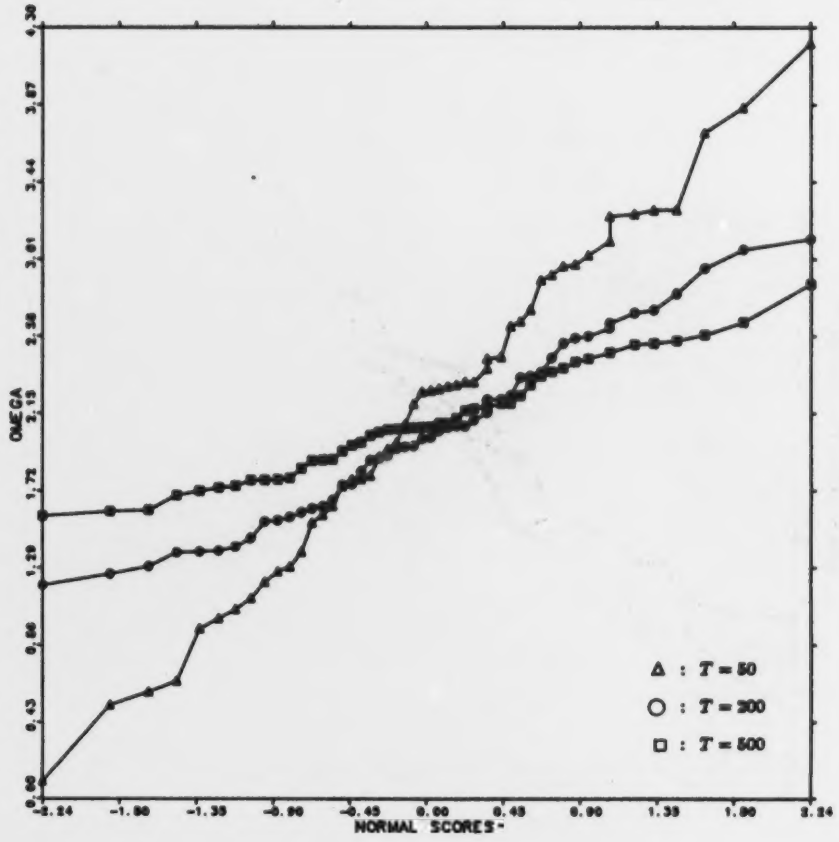


Figure 2

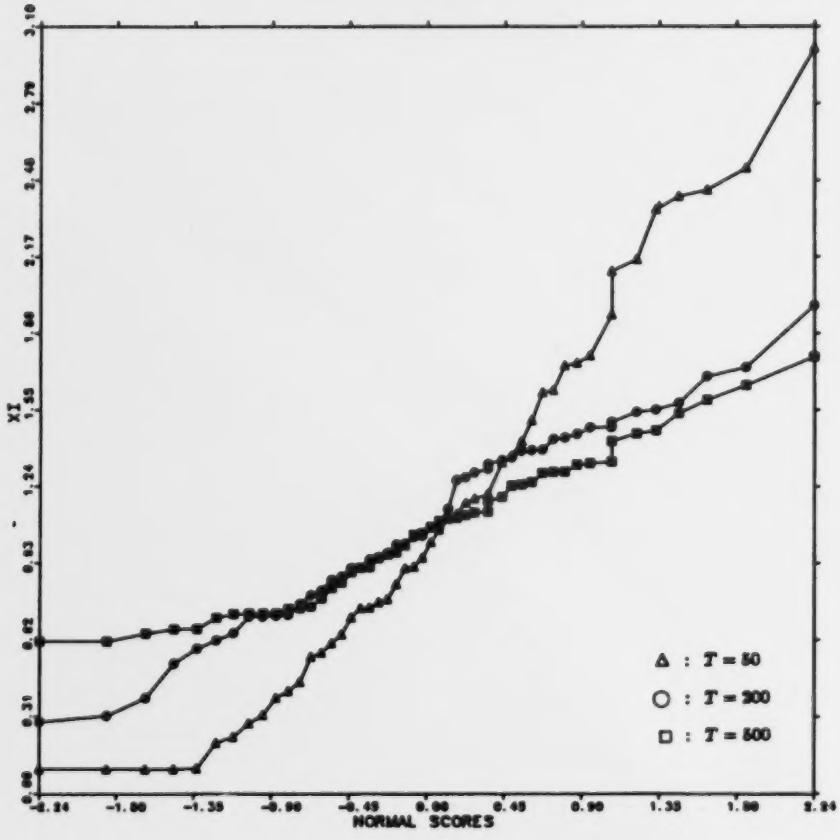


Figure 3

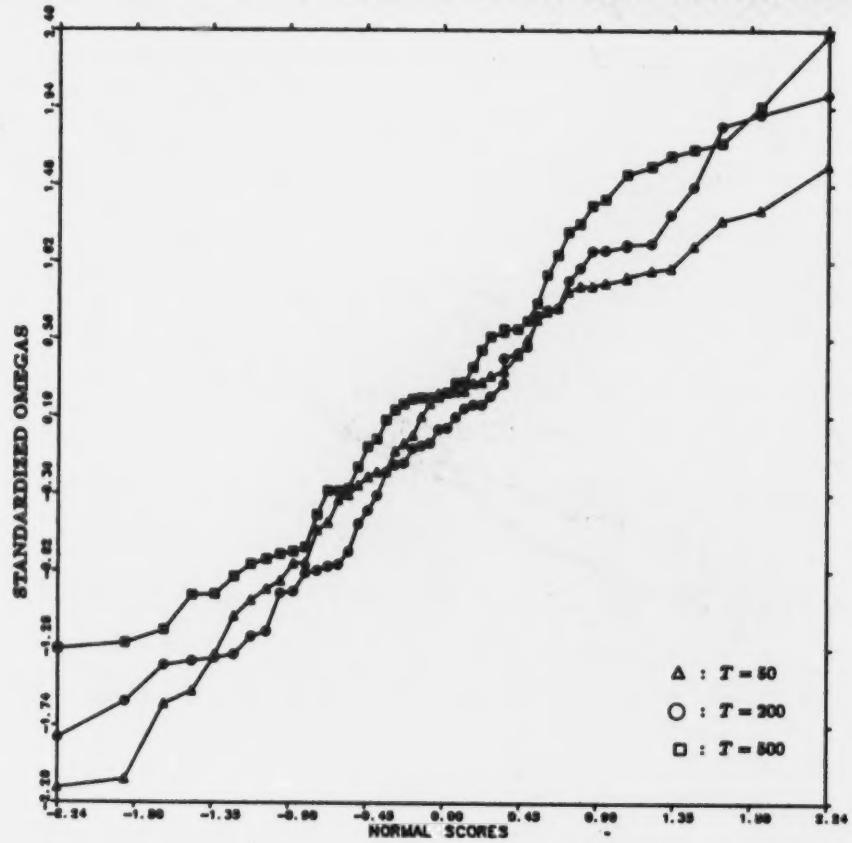
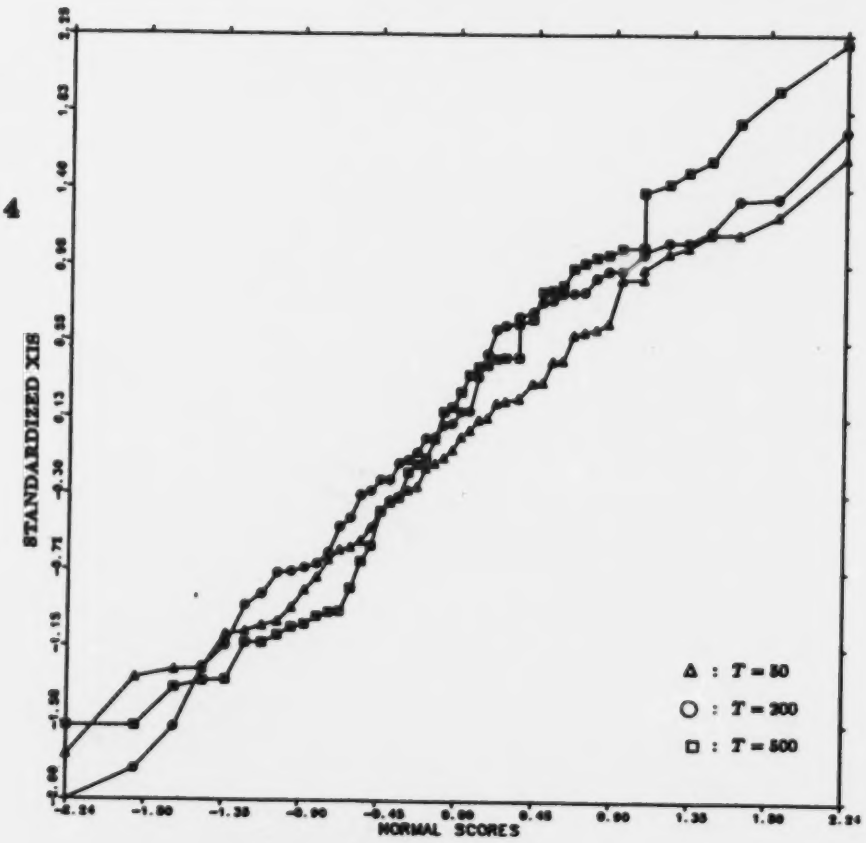


Figure 4



6. Adaptive Filtering and Smoothing with an Example

Two different types of adaptive filters can be constructed from the scoring recursions of §4. A fast filter that does not require storage of past data can be described in the following steps:

1. Save $\hat{\theta}_{t-1}$ and estimates of the Fisher information matrix (or its inverse), the score function, $\tilde{z}_{t-1}$, R_{t-1}, $\{\dot{\tilde{z}}_{t-1}\}$, and $\{\dot{R}_{t-1}\}$.
2. Observe x_t. Use $\hat{\theta}_{t-1}$ to estimate the additional terms involved with updating the last six items of step 1. Now update these items.
3. Score $\hat{\theta}_{t-1}$, a fixed number of times using the updated scoring information from step 2. Call this $\hat{\theta}_t$ and return to step 1.

Numerical runs of this filter with the model of §5 showed $\{\hat{\theta}_t\}$ to be unstable with values of $\hat{\theta}_t$ wandering about without direction. This instability is not due to "getting a bad start" since the same behaviour was observed when T was taken as 900 and the 7 items of step 1 were started with their MLE's based on the first 500 observations. The failure of this filter suggests that it may not be possible to consistently estimate θ with an adaptive filter without storing past data.

The second adaptive filter is slower, requires past data storage, and is based on the computation of a new MLE of θ after each observation. One particular simulation from §4 with $T = 200$ was used to investigate this filter. If MLE $\hat{\theta}_t$ based on $x_1, \ldots, x_t$ is computed by scoring, then it does not suffice to simply start computing $\hat{\theta}_t$ by scoring initial estimate $\hat{\theta}_{t-1}$. Such a procedure soon gets caught at $\hat{\Omega}_t = .1$ and cannot move away as more data is observed. This problem is corrected by including (for each t) an iteration sequence starting at a randomly chosen θ and using the best of the two iteration sequences in calculating $\hat{\theta}_t$. The results of such calculations using 50 steps of scoring for each $\hat{\theta}_t$ are shown in Figure 5 where $\hat{\theta}_t$ for $t = 101, \ldots, 175$ is plotted with connecting lines ($\hat{\Omega}_t$ on top and $\hat{\Xi}_t$ on the bottom). The other sequence of $\hat{\theta}_t$'s in this figure was computed using the EM algorithm allowing 100 iterations at each t. Unlike the scoring method, it was sufficient in calculating $\hat{\theta}_t$ to start iterations at $\hat{\theta}_{t-1}$ alone. For this model

the scored values of $\hat{\theta}_t$ are always greater than those from the EM algorithm. The middle plot of Figure 5 shows the real time estimates of standard deviations for $\hat{\Omega}_t$ (connected boxes) and $\hat{\Theta}_t$ (circles) that are computed like $\hat{\theta}_t$ using only present and past data. Figures 6 and 7 demonstrate the tracking ability of the adaptive filter. The solid line of Figure 6 is the true state space while the solid stars represent the real time adaptive filter $\{\tilde{z}_t = \tilde{z}_t(\hat{\theta}_t)\}$ and the circles denote filter $\{\tilde{z}_t = \tilde{z}_t(\hat{\theta}_{200})\}$ with $(\hat{\Omega}_{200}, \hat{\Theta}_{200}) = (2.0816, .9118)$. There was no graphical difference between the adaptive filter constructed from the scored version of $\{\hat{\theta}_t\}$ and the EM version so the solid stars could stand for either. In fact there is little difference between the two filters in Figure 6. Figure 7 plots real time estimates of $\Sigma_t = \mathrm{Cov}_\theta(\tilde{z}_t)$.

Figure 8 compares the adaptive filter $\{\tilde{z}_t = \tilde{z}_t(\hat{\theta}_{200})\}$ represented as circles with the adaptive smoother $\{\hat{z}_t = \hat{z}_t(\hat{\theta}_{200})\}$ given as triangles for every third value of t ranging from 34 to 175. The smoother performs a little better as one might expect since $\hat{z}_t$ is the expectation of z_t conditioning on all the data and not just past and present data as with $\tilde{z}_t$.

The most striking feature of this example is the jump of $\hat{\Omega}_t$ in Figure 5 between $t = 142$ and 143. An inspection of the top two plots of Figure 9, showing the observational $\{\epsilon_t\}$ and state space $\{f_t\}$ errors in the simulation, reveals $\epsilon_{143} = 3.64$ as an $N(0,2)$ variate at the 99.49 percentile. Outlier x_{143} is easily detected by inspecting the two residual plots at the bottom of Figure 5. The left plot is the sequence of adaptive innovations $\{x_t - A_t B_t \tilde{z}_{t-1}(\hat{\theta}_{t-1})\}$ studentized by standard deviation estimates as described in Appendix 3. The right plot is of $\{\tilde{z}_t(\hat{\theta}_t) - B_t \tilde{z}_{t-1}(\hat{\theta}_{t-1})\}$ also studentized. The outlier produces large residuals at $t = 143$ in both plots. Note that $f_{140} = 2.79$ is at the 99.74 percentile and even more extreme than ϵ_{143}. However, since it involves only unobservable data, it has no influence on our data analysis and we refrain from calling it an "outlier." Note that because of f_{140}, both filters and the smoother are far off the mark in predicting z_{140} but recover well in predicting z_{141}.

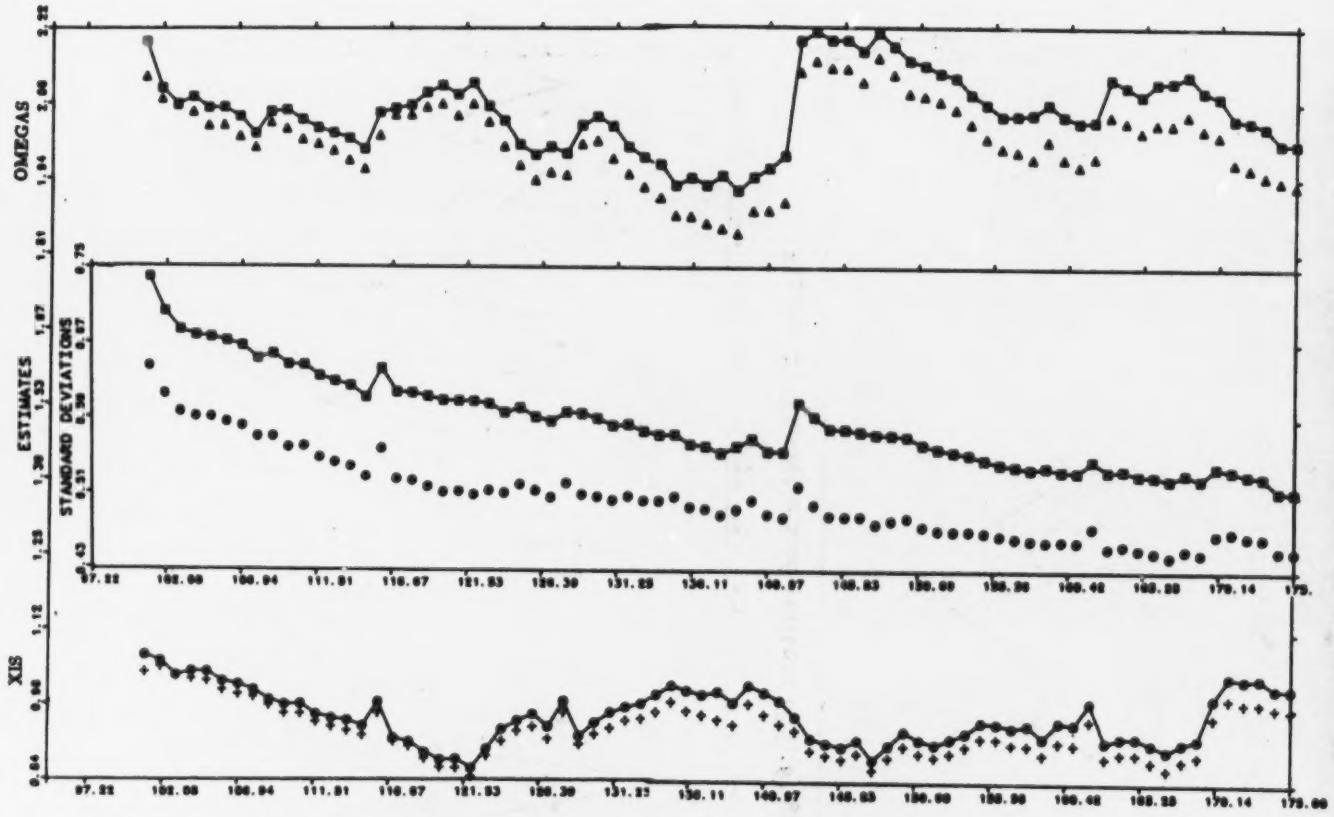


Figure 5. Plots of $(\hat{\Omega}_t, \hat{\Xi}_t)$ vs. t for the method of scoring (solid lines) and the EM algorithm. The middle plot shows the corresponding real time estimates of standard deviations for $\hat{\Omega}_t$ (connected boxes) and $\hat{\Xi}_t$ (circles) derived from the inverse Fisher information estimate.

Figure 6. Tracking the state space (boxes) with adaptive filters $\{\tilde{x}_t = \tilde{x}_t(\hat{\theta}_t)\}$ (solid stars) and $\{\tilde{x}_t = \tilde{x}_t(\hat{\theta}_{200})\}$ (circles).

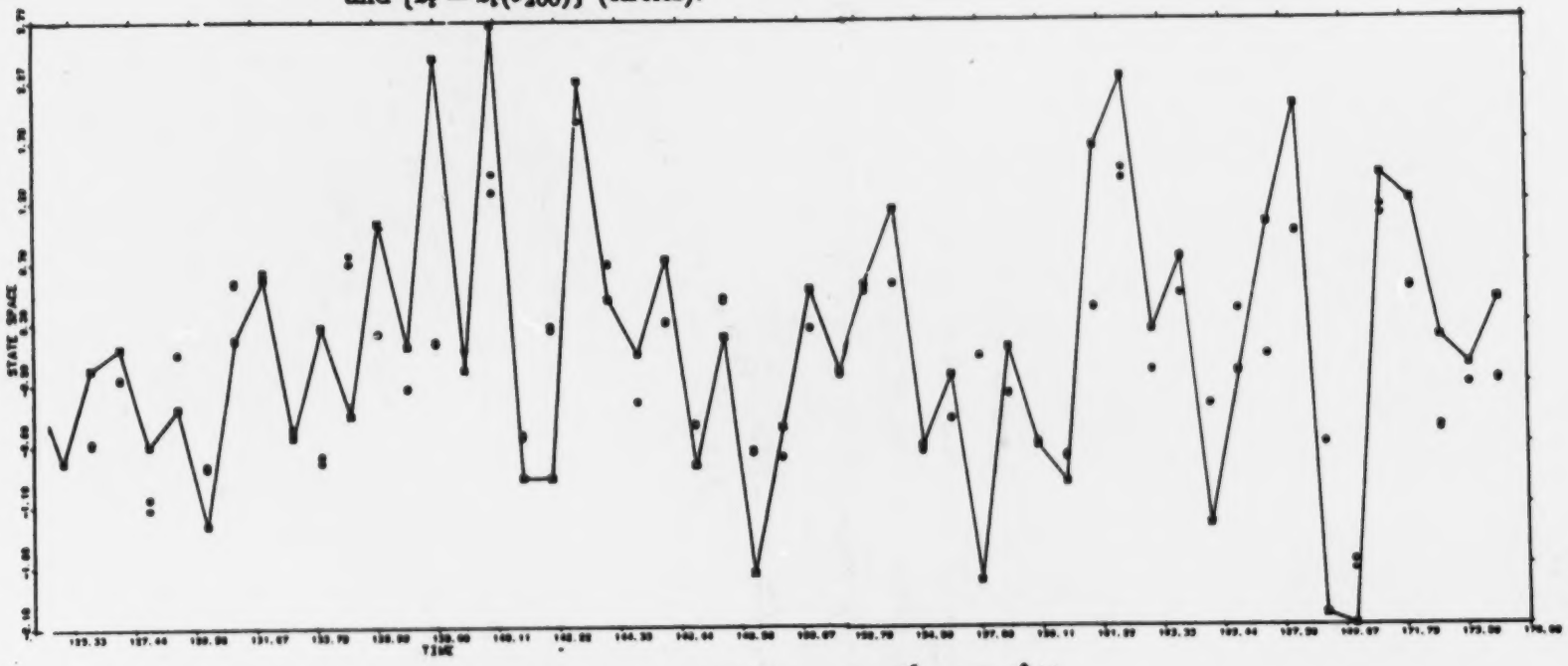
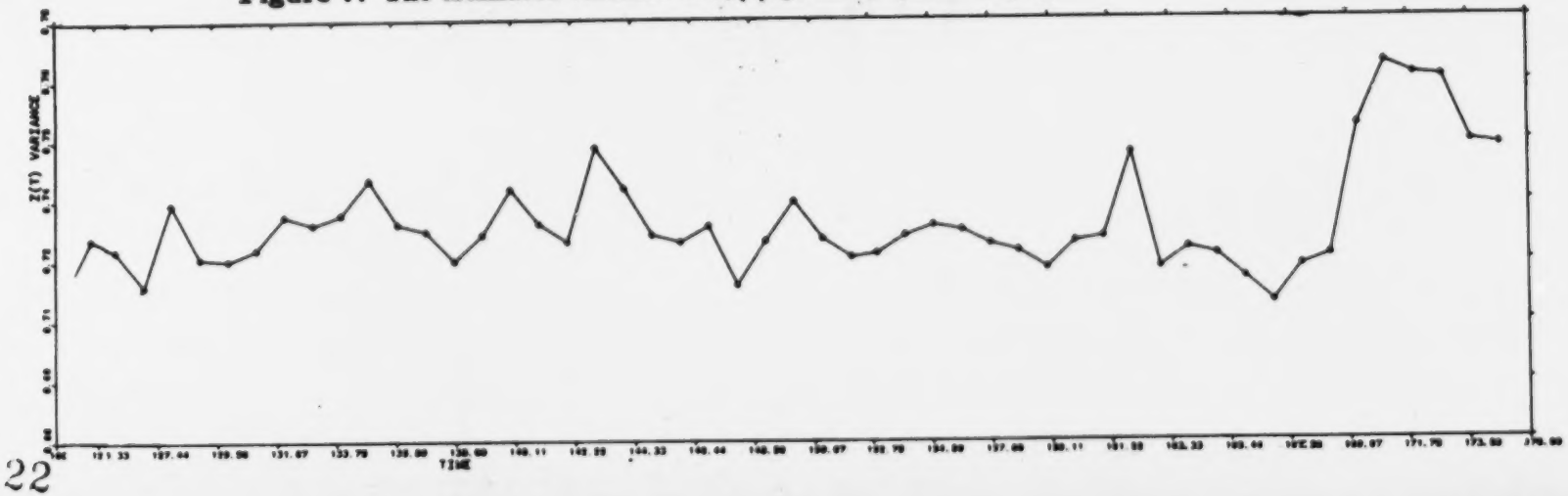


Figure 7. The estimated variance of $\tilde{x}_t(\theta)$ given by $\{\hat{\Sigma}_t = \Sigma_t(\hat{\theta}_t)\}$.



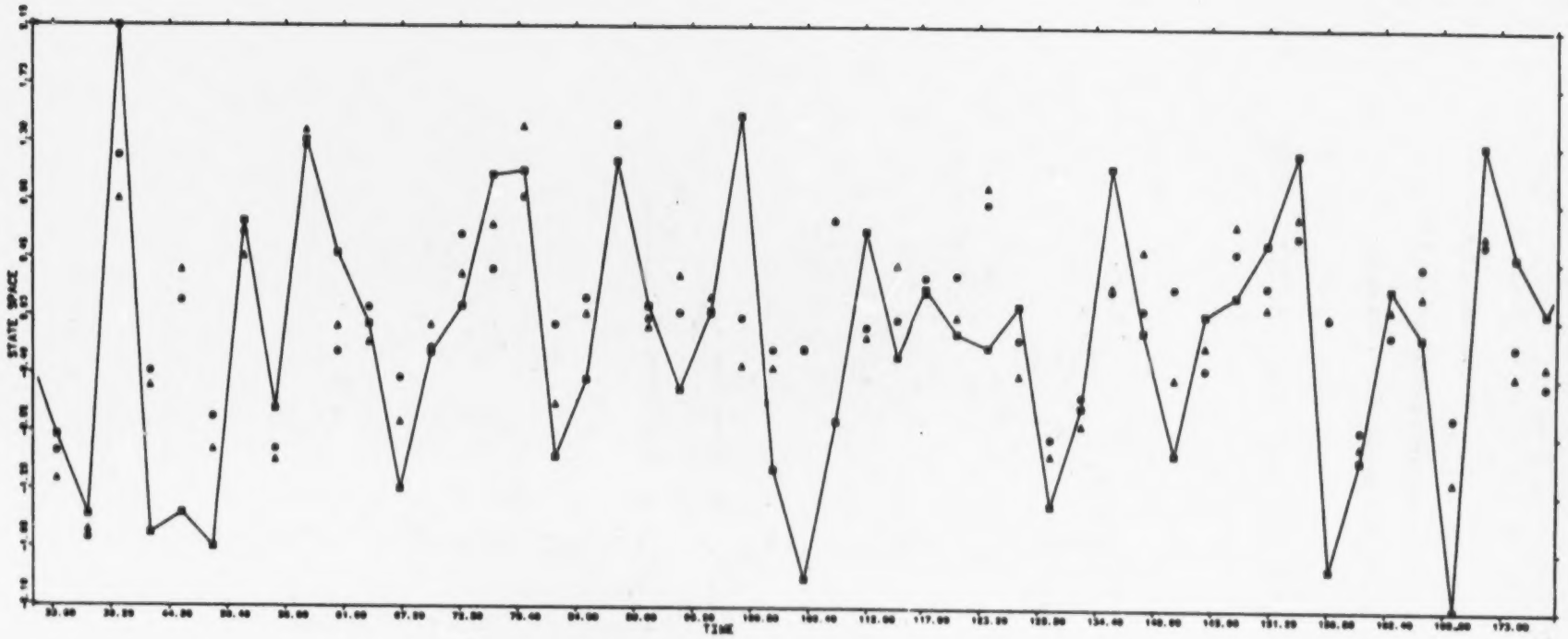


Figure 8. A comparison of the adaptive filter $\{\tilde{x}_t(\hat{\theta}_{200})\}$ represented as circles with adaptive smoother $\{\hat{x}_t(\hat{\theta}_{200})\}$ given by triangles. The true state space follows the connected boxes.

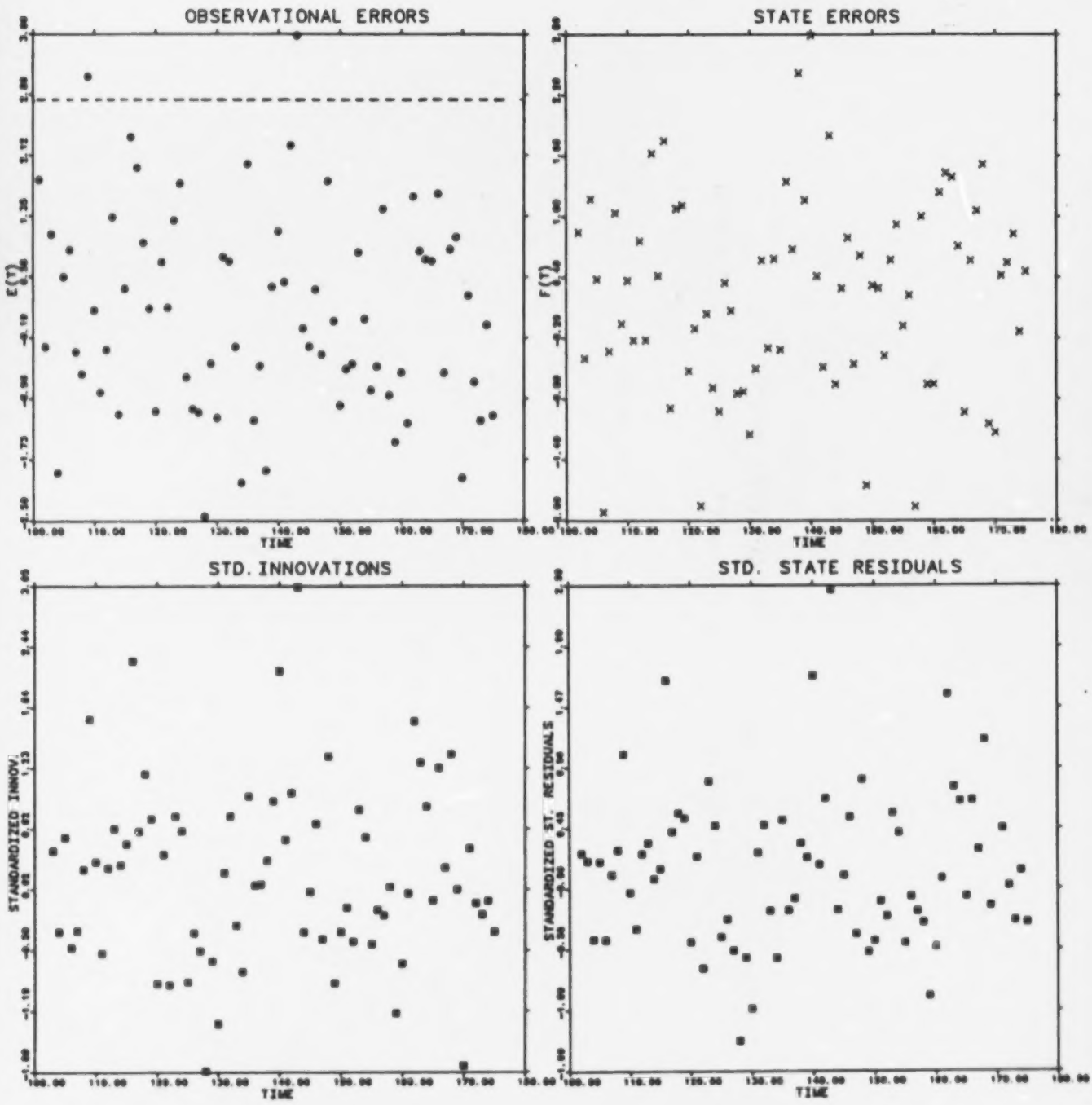


Figure 9. The top plots are comprised of unobservable modelling errors while the bottom residual plots are based entirely on the data. The dotted line marks 2 standard deviations from zero.

Appendix 1

Backward recursions for the smoothing equations of (3.4) are derived under the condition that the filtering values $\{\tilde{z}_t, \Sigma_t, R_t\}$ are available. Following Jaswinski (1970, p. 216-218), the posterior of z_t and z_{t+1} given all the data x and θ fixed is

$$f(z_{t+1}, z_t \mid x) \propto f(z_{t+1}, z_t, x)$$
$$\propto f(z_{t+1}, \ldots, z_T \mid z_{t+1}, z_t, X_t) f(z_{t+1} \mid z_t, X_t) f(z_t \mid X_t)$$

where $X_t = (x_1, \ldots, x_t)$. From (1.1) and (1.2) then

$$f(z_{t+1}, z_t \mid x) \propto f(z_{t+1}, \ldots, z_T \mid z_{t+1}) f(z_{t+1} \mid z_t) f(z_t \mid X_t). \qquad (A1.1)$$

Each of the three factors in (A1.1) is a normal density so the joint posterior is bivariate normal. The value $\hat{z}_t$ can be found from $\hat{z}_{t+1}$ by replacing z_{t+1} with $\hat{z}_{t+1}$ in (A1.1) and maximising the last two factors over z_t. Since

$$z_{t+1} \mid z_t \sim MVN_r(B_{t+1} z_t, \Xi) \quad \text{and} \quad z_t \mid X_t \sim MVN_r(\tilde{z}_t, \Sigma_t)$$

then

$$\hat{z}_t = (\Sigma_t^{-1} + B_{t+1}' \Xi^{-1} B_{t+1})^{-1} (\Sigma_t^{-1} \tilde{z}_t + B_{t+1}' \Xi^{-1} \hat{z}_{t+1}).$$

Using the identities

$$(\Psi + VQW)^{-1} = \Psi^{-1} - \Psi^{-1} V (Q^{-1} + W \Psi^{-1} V)^{-1} W \Psi^{-1} \qquad (A1.2)$$

and

$$(\Psi + VQW)^{-1} VQ = \Psi^{-1} V (W \Psi^{-1} V + Q^{-1})^{-1} \qquad (A1.3)$$

then the recursion of $\hat{z}_t$ from $\hat{z}_{t+1}$ follows. The posterior precision matrix is determined by considering the quadratic terms contributed by each of the factors in (A1.1). The quadratics involved lead to $\{\mathrm{Var}(z_{t+1}', z_t' \mid x)\}^{-1}$ as

$$\begin{pmatrix} \Xi^{-1} + W & -\Xi^{-1} B_{t+1} \\ -B_{t+1}' \Xi^{-1} & \Sigma_t^{-1} + B_{t+1}' \Xi^{-1} B_{t+1} \end{pmatrix} = \begin{pmatrix} V_{t+1} & V_{t+1,t} \\ V_{t,t+1} & V_t \end{pmatrix}^{-1} \qquad (A1.4)$$

where W represents the unknown contribution of the first factor in (A1.1) to the precision of z_{t+1}. The unknowns in (A1.4) include W, V_t, and $V_{t+1,t}$ while V_{t+1} is

assumed known. This equation can be solved for each unknown using the inversion identity for a partitioned matrix (see Press (1971, §2.6.1). Then

$$(\Xi^{-1} + W - \Xi^{-1}B_{t+1}(\Sigma_t^{-1} + B_{t+1}'\Xi^{-1}B_{t+1})^{-1}B_{t+1}'\Xi)^{-1} = V_{t+1}$$

and after applying (A1.2) with $\Psi = \Xi$ then

$$(W + (\Xi + B_{t+1}\Sigma_t B_{t+1}')^{-1})^{-1} = V_{t+1}$$

so that $W = V_{t+1}^{-1} - R_{t+1}^{-1}$. Note that W represents the difference in the precision of the smoothed estimator and the one-step ahead predictor. Having solved for W, the left side of (A1.4) is entirely known and the recursion for $V_{t+1,t}$ in (3.4) is straightforward. A computationally more efficient form of this recursion results by noting that

$$\Xi^{-1} - R_{t+1} = \Xi^{-1}B_{t+1}Q^{-1}B_{t+1}'\Xi^{-1}, \qquad (A1.5)$$

with $Q = \Sigma_t^{-1} + B_{t+1}'\Xi^{-1}B_{t+1}$, which follows from inversion of R_{t+1} with (A1.2). Then

$$V_{t+1,t} = (V_{t+1}^{-1} + \Xi^{-1}B_{t+1}Q^{-1}B_{t+1}'\Xi^{-1})^{-1}\Xi^{-1}B_{t+1}V_t$$

and the first factor can also be inverted as in (A1.2) with $\Psi = V_{t+1}^{-1}$ so that

$$\begin{aligned} V_{t+1,t} = {} & V_{t+1}\Xi^{-1}B_{t+1}V_t \\ & - V_{t+1}\Xi^{-1}B_{t+1}\Sigma_t(\Sigma_t + \Sigma_t B_{t+1}'\Xi^{-1}(\Xi + V_{t+1})\Xi^{-1}B_{t+1}\Sigma_t)^{-1} \\ & \cdot \Sigma_t B_{t+1}'\Xi^{-1}V_{t+1}\Xi^{-1}B_{t+1}V_t, \end{aligned}$$

which involves the inversion of one matrix for each t instead of three as in (3.4).

The recursion for V_t follows from solving (A1.4) with the inversion identity for a partitioned matrix. Then

$$V_t = (\Sigma_t^{-1} + B_{t+1}'\Xi^{-1}B_{t+1} - B_{t+1}'\Xi^{-1}(\Xi^{-1} + V_{t+1}^{-1} - R_{t+1}^{-1})^{-1}\Xi^{-1}B_{t+1})^{-1}$$

which can be reduced by substituting for Q and $\Xi^{-1} - R_{t+1}^{-1}$ as in (A1.5) so that

$$V_t = (Q - B_{t+1}'\Xi^{-1}(V_{t+1}^{-1} + \Xi^{-1}B_{t+1}Q^{-1}B_{t+1}'\Xi^{-1})^{-1}\Xi^{-1}B_{t+1})^{-1}.$$

Note that QV_tQ assumes the form (A1.2) with $\Psi = Q$ so that

$$\begin{aligned} V_t &= Q^{-1}(Q + B'_{t+1}\Xi^{-1}V_{t+1}\Xi^{-1}B_{t+1})Q^{-1} \\ &= Q^{-1} + (\Sigma_t^{-1} + B'_{t+1}\Xi^{-1}B_{t+1})^{-1}B'_{t+1}\Xi^{-1}V_{t+1}\Xi^{-1}B_{t+1}Q^{-1}. \end{aligned}$$

Expansion of the first Q as in (A1.2) and two applications of (A1.3) to the second factor leads to the desired result in (3.4).

Appendix 2

The distribution $f(x \mid \theta)$ is MVN so that the sequence

$$f(x_1 \mid \theta), f(x_2 \mid x_1, \theta), \ldots, f(x_T \mid x_1, \ldots, x_{T-1}, \theta)$$

also consists of normal densities. Therefore the essential portion of $\ell n\, f(x \mid \theta) = \ell$ is

$$\ell = -1/2 \sum_{t=1}^{T} \{\ell n|G_t| + \text{tr } G_t^{-1}(x_t - m_t)^2\} \qquad (A2.1)$$

where $m_t = E(x_t \mid X_{t-1}, \theta)$ with $X_{t-1} = (x_1, \ldots, x_{t-1})$ and

$$G_t = \text{Cov}\{x_t \mid X_{t-1}, \theta\}.$$

To find m_t note that

$$\begin{aligned} m_t &= E_\theta(A_t z_t + \epsilon_t \mid X_{t-1}) = A_t E_\theta(z_t \mid X_{t-1}) \\ &= A_t E_\theta(B_t z_{t-1} + f_t \mid X_{t-1}) = A_t B_t \tilde{z}_{t-1}. \end{aligned} \qquad (A2.2)$$

Also

$$\begin{aligned} G_t &= \text{Cov}_\theta\{A_t z_t + \epsilon_t \mid X_{t-1}\} = A_t \text{Cov}_\theta\{z_t \mid X_{t-1}\}A'_t + \Omega \\ &= A_t R_t A'_t + \Omega. \end{aligned} \qquad (A2.3)$$

Expressions (A2.1)-(A2.3) make it possible to derive the recursive formulae of (4.3) and (4.6) for calculating Fisher information. Before proceeding, however, it is convenient to summarize a few results from matrix calculus.

Lemma 1. *Suppose G and H are $(k \times k)$ and symmetric and each component of G and H is a function of scalar θ. Then*

$$\text{(a)} \qquad \partial|G|/\partial\theta = |G|\text{tr } G^{-1}\dot{G} \quad \text{where} \quad \dot{G} = \partial G/\partial\theta \qquad (A2.4)$$

$$\text{(b)} \qquad \partial\text{tr } G^{-1}H/\partial\theta = -\text{tr } G^{-1}HG^{-1}\dot{G} + \text{tr } G^{-1}\dot{H}. \qquad (A2.5)$$

Proof. The results follow from a patient application of the chain rule and product rule in combination with Mardia, Kent, and Bibby (1979, equations (A.9.3)-(A.9.5)).

In order to compute $I(\omega_{ij}^2) = -E\{\partial^2\ell/\partial\omega_{ij}^2\} = -E(\bar{\ell})$, we differentiate (A2.1) using (A2.4) and (A2.5). Then, applying the chain rule via $|G_t|, x_t - m_t$, and G_t^{-1},

$$-2\dot{\ell} = \sum_{t=1}^{T}\{\text{tr } G_t^{-1}\dot{G}_t + 2(m_t - x_t)'G_t^{-1}\dot{m}_t - \text{tr } G_t^{-1}(m_t - x_t)^2G_t^{-1}\dot{G}_t\}. \qquad (A2.6)$$

In computing $\bar{\ell}$ we must again apply the chain rule via $G_t^{-1}, \dot{G}_t, m_t - x_t$, and $\dot{m}_t$, but can ignore terms resulting in $m_t - x_t$ since its expectation is zero. This leads to

$$\begin{aligned}-2\bar{\ell} = \sum_t\{&-\text{tr } G_t^{-1}\dot{G}_tG_t^{-1}\dot{G}_t + \text{tr } G_t^{-1}\bar{G}_t + 2\dot{m}_t'G_t^{-1}\dot{m}_t \\ &+ \text{tr } G_t^{-1}(m_t - x_t)^2G_t^{-1}\dot{G}_tG_t^{-1}\dot{G}_t + \text{tr } G_t^{-1}\dot{G}_tG_t^{-1}(m_t - x_t)^2G_t^{-1}\dot{G}_t \\ &- \text{tr } G_t^{-1}(m_t - x_t)^2G_t^{-1}\bar{G}_t\}.\end{aligned} \qquad (A2.7)$$

By taking the expectation with $E(m_t - x_t)^2 = G_t$, then

$$I(\omega_{ij}^2) = 1/2\sum_{t=1}^{T}\{\text{tr } G_t^{-1}\dot{G}_tG_t^{-1}\dot{G}_t + 2\dot{m}_t'G_t^{-1}\dot{m}_t\}. \qquad (A2.8)$$

In (A2.8) all dots refer to $\partial/\partial\omega_{ij}$. Expressions for $I(\xi_{ij}^2)$ and $I(\omega_{ij}\xi_{k\ell})$ have the same form with all the dots based on $\partial/\partial\xi_{ij}$ in the former and one dot in each term for $\partial/\partial\omega_{ij}$ and $\partial/\partial\xi_{k\ell}$ in the latter. Since tr AB = vec$(A)'$vec(B) for A and B symmetric, then by defining

$$h^t = \text{vec}(G_t^{-1}\partial G_t/\partial\omega_{ij}G_t^{-1}) \quad \text{and} \quad f^t = \text{vec}(\partial G_t/\partial\xi_{k\ell})$$

we may write (A2.8) as

$$I(\omega_{ij}\xi_{k\ell}) = \tfrac{1}{2}\sum_{t=1}^{T}\{(h^t)'f^t + 2(\partial m_t/\partial\omega_{ij})'G_t^{-1}\partial m_t/\partial\xi_{k\ell}\}. \qquad (A2.9)$$

Expression (4.3) follows from the form of (A2.9).

Proofs of the recursions in (4.7)-(4.10) are now given. The following lemma will be needed.

Lemma 2. *Suppose G and H are symmetric matrices whose components depend on scalar θ. Then*

$$\partial G^{-1}H/\partial\theta = G^{-1}\dot{H} - G^{-1}\dot{G}G^{-1}H. \qquad (A2.10)$$

Proof. Set $G^{-1}H = L$ and write as $H = GL$. Differentiating this expression and solving for $\dot{L}$ gives (A2.10).

The expression for $\dot{\Sigma}_t$ in (4.7) follows from a straightforward differentiation of (2.4) using the product rule and the lemma above. Expression (4.8) for $\dot{\tilde{z}}_t$ also follows from differentiating (2.4). Differentiation gives

$$\begin{aligned}\dot{\tilde{z}}_t = {} & (I - R_tA_t'G_t^{-1}A_t)B_t\,\dot{\tilde{z}}_{t-1} \\ & + [\dot{R}_tA_t' - R_tA_t'G_t^{-1}(A_t\dot{R}_tA_t' + \dot{\Omega})]G_t^{-1}(x_t - A_tB_t\tilde{z}_{t-1}).\end{aligned}$$

Using (A1.3) on $R_tA_t'G_t^{-1}$ in both terms and reducing gives

$$\begin{aligned}\dot{\tilde{z}}_t &= (A_t'\Omega^{-1}A_t + R_t^{-1})^{-1}R_t^{-1}B_t\,\dot{\tilde{z}}_{t-1} \\ &\quad + [\dot{R}_tA_t' - (A_t'\Omega^{-1}A_t + R_t^{-1})^{-1}A_t'\Omega^{-1}(A_t\dot{R}_tA_t' + \dot{\Omega})]G_t^{-1}(x_t - A_tB_t\tilde{z}_{t-1}) \\ &= (A_t'\Omega^{-1}A_t + R_t^{-1})^{-1}[R_t^{-1}B_t\,\dot{\tilde{z}}_{t-1} \\ &\quad + (R_t^{-1}\dot{R}_tA_t' - A_t'\Omega^{-1}\dot{\Omega})G_t^{-1}(x_t - A_tB_t\tilde{z}_{t-1})].\end{aligned}$$

Applying (A1.2) to the first factor gives (4.8). Expression (4.9) follows in a straightforward way and (4.10) follows the argument used to derive (4.8).

Appendix 3

For θ fixed the covariance of innovation $i_t = x_t - A_t B_t \tilde{z}_{t-1}$ is computed by substituting $x_t = A_t z_t + \epsilon_t$ and using $\mathrm{Var}_\theta\{z_{t-1} - \tilde{z}_{t-1}\} = \Sigma_{t-1}$ to show that

$$\mathrm{Cov}_\theta(i_t) = \Omega + A_t R_t A_t' = G_t = \mathrm{Cov}_\theta(x_t \mid x_1, \ldots, x_{t-1}).$$

The suggested studentization of $\hat{i}_t = i_t(\hat{\theta}_t)$ is therefore $\hat{i}_t' \hat{G}_t^{-1}(\hat{\theta}_t) \hat{i}_t$.

For state space residuals consider the covariance of $\tilde{f}_t = \tilde{z}_t - B_t \tilde{z}_{t-1}$ which follows from (2.4) as

$$\begin{aligned} \mathrm{Cov}_\theta(\tilde{f}_t) &= R_t A_t' G_t^{-1} \, \mathrm{Cov}(i_t) G_t^{-1} A_t R_t' \\ &= R_t A_t' G_t^{-1} A_t R_t' = R_t - \Sigma_t. \end{aligned}$$

The suggested studentization of $\hat{f}_t = \tilde{z}_t(\hat{\theta}_t) - B_t \tilde{z}_{t-1}(\hat{\theta}_{t-1})$ is $\hat{f}_t' (R_t(\hat{\theta}_t) - \Sigma_t(\hat{\theta}_t))^{-1} \hat{f}_t$.

Acknowledgment. Thanks to Marco Tucci for many interesting discussions.

References

1. Butler, R.W. (1985). Predictive likelihood inference with applications (with discussion). J. Roy. Statist. Soc. B, to appear.
2. Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Statist. Soc. B **39**, 1-38.
3. Harrison, P.J. and Stevens, C.F. (1971). A Bayesian approach to short-term forecasting. Oper. Res. Quarterly **22**, 341-362.
4. ——————(1976). Bayesian forecasting (with discussion). J. Roy. Statist. Soc. B **38**, 205-247.
5. Jazwinski, A.H. (1970). Stochastic Processes and Filtering Theory. New York: Academic Press.
6. Kendall, M.G. and Stuart, A. (1973). The Advanced Theory of Statistics, vol. **2**. New York: Hafner.
7. Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). J. Roy. Statist. Soc. B **34**, 1-41.
8. Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). Multivariate Analysis. London: Academic Press.
9. Meinhold, R.J. and Singpurwalla, N.D. (1983). Understanding the Kalman filter. The Amer. Statistician **37**, 123-127.
10. Press, S.J. (1971). Applied Multivariate Analysis. New York: Holt, Rinehart, and Winston.